

Bayesian estimation and regularization techniques in categorical data analysis

J. KALINA

Abstract

This paper explores Bayesian estimation for categorical data, focusing on simple yet effective models that provide a foundation for applying more advanced methods accurately and reliably in real-world applications. We begin by revisiting Bayesian estimators for the binomial distribution and investigating their properties. Next, we develop hypothesis tests for categorical data (sign test, homogeneity test, symmetry test) based on regularized maximum likelihood estimates of the probabilities. Finally, we formulate regularized versions of common association measures for contingency tables and study the regularized version of mutual information, particular for the situation where the regularized version can effectively handle zero counts.

Mathematics Subject Classification 2000: 62H17, 62F15

General Terms: Categorical data, Bayesian estimation, Hypothesis tests, Regularization

Keywords: contingency tables, mutual information, zero-count adjustment, homogeneity test, symmetry test.

1. INTRODUCTION

Bayesian thinking, with its origins in [Bayes 1763], has recently penetrated many applications in various fields. Bayesian estimation is an important approach in the analysis of categorical data [Johnson et al. 2022], particularly for contingency tables (two-way frequency tables) that aggregate counts according to different categorical variables [Agresti 2002]. Analyzing complex Bayesian models requires understanding the basic principles of simpler models [Loftus 2024], as well as the principles for selecting appropriate prior distributions [Tuyl et al. 2008].

This motivates us to revisit the binomial distribution and two-way contingency tables as simple models for categorical data, and to consider Bayesian estimation of their parameters along with regularized versions of hypothesis tests or association measures. The broader aim is to develop tailor-made tools for reliable analysis of complex categorical data, increasingly relevant in modern applications [Lindskou et al. 2020]. Such methods are also relevant for machine learning, where regularized

This work is supported by the project 24-11146S (“Maximal entropy portfolio”) of the Czech Science Foundation.

10.2478/jamsi-2025-0011

©J. Kalina

This is an open access article licensed under the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>).

Bayesian estimation can improve the accuracy and robustness of predictions, similarly to how neural networks learn complex patterns [Wang et al. 2024].

Common types of regularized estimators are known to be connected to Bayesian thinking [Jaynes 2003]. The recent boom of regularized estimators has focused mainly on continuous data, often aimed at sparse methods for variable selection in classification [Ledoit and Wolf 2022] or covariance matrix estimation [Kalina and Tichavský 2022].

For categorical data, the motivation for regularization may be quite different. Regularization can improve robustness to mismatches, i.e., situations when a measurement is erroneously assigned to the wrong category (e.g., confusing success and failure for binary data). It also reduces the variability of estimates, because maximum likelihood estimates have too high a variability for small samples; introducing a small bias can remarkably reduce this variability [Sohae 2023]. Moreover, regularization directly helps mitigate numerical instability in test statistics, parameter estimates, association measures, and mutual information, improving the reliability and robustness of inference. Regularization is particularly useful when some counts in the contingency table are very small (or even 0), especially if the number of categories of one or more variables is large. While this paper does not address such high-dimensional settings, they represent an important direction for future research.

Regularized estimators often take the form of shrinkage estimators. For Bayesian tools, it is natural to obtain the shrinkage estimators as shrunken versions of the maximum likelihood estimator (MLE) towards the mean of the prior distribution. If the categorical data are in several different groups, then the shrinkage is natural to be considered towards the MLE evaluated across the groups. Recent proposals of regularized methods for categorical data include hypothesis tests [Wang and Li 2023], clustering [Baek and Park 2023], logistic regression models [Ming and Yang 2024], or monitoring categorical processes [Wang et al. 2023]. The connection of regularization and Bayesian estimation has been investigated for continuous data in [Zhou et al. 2024]. Regularized tools (mainly test statistics) for analyzing categorical data have been used for the labor market segmentation in [de Toledo et al. 2020]. Much more intensive attention has been paid to regularized test statistics for continuous data, e.g. for the Hotelling's T^2 test [Issouani et al. 2024].

In current machine learning, pattern discovery or association rule learning represents an important task [Golden 2020]. So far, individual applications exploiting

regularized approaches have considered mainly continuous data. For example, the learned patterns have been used to construct classification rules or to search for latent clusters [Braun et al. 2017]. Rare examples of pattern discovery on categorical data are cited in the overview [Subramanian et al. 2020] for the integration of genomics or metabolomics data. It is therefore natural to assume that effective pattern discovery in contingency tables should rely on suitable regularized estimation tools, as these are able to stabilize inference in the presence of sparse counts and improve the reliability of detected associations.

The concept of entropy plays a central role in modern quantitative finance, where it underpins portfolio construction methods designed to remain stable under model uncertainty. In this context, maximal entropy approaches aim to distribute weights as evenly as possible while respecting structural constraints, thus promoting diversification and robustness. A similar rationale applies in the analysis of categorical data: when faced with sparse or zero counts, entropy-based reasoning naturally leads to regularized estimators that stabilize inference by shrinking toward more balanced distributions. Entropy provides a conceptual link between applications in finance and the analysis of categorical data. In both cases, entropy-based reasoning encourages balanced and robust estimates, which is particularly valuable when counts are sparse or zero [Gupta et al. 2025].

This paper presents a methodological application of regularized Bayesian estimation for categorical data, focusing on handling challenges such as small or zero counts. Section 2 is devoted to the estimation of the parameter of the binomial distribution. There, regularized estimators obtained as Bayesian tools with different choices of the prior distribution are studied and discussed. Section 3 studies commonly used hypothesis tests for categorical data for the situations with regularized maximum likelihood estimates of probabilities for individual cells. We formulate regularized versions of some common association measures for contingency tables in Section 4 and study the regularized version of mutual information in Section 5. Section 6 concludes the paper.

2. BERNOULLI DISTRIBUTION

Let us consider i.i.d. random variables $X_1, \dots, X_n$ coming from the Bernoulli distribution (alternative distribution, 0-1 distribution) with an unknown parameter $\pi \in [0, 1]$. In this section, we recall several well-known versions of the Bayesian estimator of π . These estimators have the form of regularized (shrinkage, penalized)

versions of the maximum likelihood estimator. We provide a novel and rigorous formal treatment of these estimators in Section 2.1.

It is common for the binomial distribution to perceive π as the probability of success and $1 - \pi$ as the probability of failure. The sum $\sum_{i=1}^n X_i$ follows the binomial distribution $\text{Bi}(n, p)$. The most prominent Bayesian estimators of π will be discussed and compared here. The MLE obtained as $\hat{\pi} = \frac{1}{n} \sum_{i=1}^n X_i$ is a consistent and unbiased estimator of π [Rao 2002]. From the theory of maximum likelihood estimation, it follows that $\hat{\pi}$ achieves the Rao-Cramér lower bound, however only under the assumption of $n \rightarrow \infty$. In a variety of applications, the MLE is not a holy grail and it may be convenient to deviate from the MLE especially for data with a small n .

Bayesian estimators of π typically have the form of regularized estimators as convex linear combinations combining the MLE with some prior information according to

$$\tilde{\pi} = \lambda \frac{\sum_{i=1}^n X_i}{n} + (1 - \lambda) E\pi, \quad \lambda \in [0, 1], \quad (1)$$

where $E\pi$ is the expectation of the prior distribution of π .

When cell counts are small, simple continuity corrections such as adding 0.5 are sometimes used to stabilize estimates. In this paper, we adopt a Bayesian approach that generalizes this primitive method: by introducing a prior distribution, the posterior mean effectively shrinks the cell probabilities in a principled way, providing improved variance reduction and numerical stability while retaining interpretability. This offers a flexible framework that extends the idea of ad hoc continuity corrections to a fully probabilistic setting.

Throughout the paper, the regularization parameter λ is considered fixed. This is in contrast to certain Bayesian choices, where the corresponding λ may implicitly depend on n (typically with $\lambda_n \rightarrow 1$ as $n \rightarrow \infty$). Our asymptotic results should therefore be interpreted under the assumption of a fixed λ . Throughout the paper, we consider the Bayesian estimates to have the form of the mean of the posterior distribution (if not stated otherwise).

Replacing $\hat{\pi}$ by a biased estimate is especially appealing if $\hat{\pi}$ is quite unexpectedly near 0 or 1, even if there is no evidence (prior belief, expectation) that it should be so. Also, the set of all possible values of the MLE, i.e. $\{0, 1/n, \dots, (n-1)/n, 1\}$, does not allow for a more delicate estimation of π for a small n , which may justify to consider small deviations from $\hat{\pi}$. Thus, we can say that the shrinkage towards $E\pi$ makes a correction for the finite sample estimation. This is especially true when $\hat{\pi}$ attains some extreme value (near 0 or 1), because the shrinkage does not change anything for

Table I. Estimators of π in the binomial model: the MLE and three Bayesian versions.

Estimator	Prior	Formula	Shrinkage (1)	
MLE	-	$\hat{\pi} = \frac{1}{n} \sum_{i=1}^n X_i$	$\lambda = 1$	-
$\tilde{\pi}_\beta$	$\text{Beta}(a, b)$	$\tilde{\pi}_\beta = \frac{\sum_{i=1}^n X_i + a}{n + a + b}$	$\lambda = \frac{n}{n + a + b}$	$E\pi = \frac{a}{a + b}$
$\tilde{\pi}_{BL}$	$U(0, 1) = \text{Beta}(1, 1)$	$\tilde{\pi}_{BL} = \frac{\sum_{i=1}^n X_i + 1}{n + 2}$	$\lambda = \frac{n}{n + 2}$	$E\pi = 1/2$
$\tilde{\pi}_J$	$\text{Beta}(1/2, 1/2)$	$\tilde{\pi}_J = \frac{\sum_{i=1}^n X_i + 1/2}{n + 1}$	$\lambda = \frac{n}{n + 1}$	$E\pi = 1/2$

$\hat{\pi} = E\pi.$

Table I overviews some useful information for three versions of the Bayesian estimator of π , i.e. for 3 different choices of the prior distribution of π , and compares them with $\hat{\pi}$. All three of them have the form (1) and the table evaluates the corresponding values of λ and $E\pi$.

- (1) **Beta prior.** First, let us consider the prior distribution to be $\text{Beta}(a, b)$. Because the two parameters $a > 0$ and $b > 0$ have to be specified, this prior represents an example of an informative prior. The estimator corresponds to the MLE obtained with $a + b$ additional experiments, where the success was achieved in a situations.
- (2) **Uniform prior.** The prior distribution $\pi \sim U(0, 1)$ coincides with $\text{Beta}(1, 1)$ and is non-informative as it does not depend on any additional parameter. As suggested by [Tuyl et al. 2008], the prior originally proposed in [Bayes 1763] is known as the Bayes-Laplace prior [Pose et al. 2021]. The uniform distribution has the largest entropy among all continuous distributions on $[0, 1]$ and using it corresponds to the max-entropy principle [Deng and Deng 2022], i.e. an important paradigm for the construction of non-informative priors.
- (3) **Jeffreys prior.** The prior distribution $\pi \sim \text{Beta}(1/2, 1/2)$ is the Jeffreys non-informative prior obtained as the square root of the Fisher information of π , which is equal to $I(\pi) = (\pi(1 - \pi))^{-1}$.

For the beta prior with parameters a and b , choosing values below 1 (i.e. $a < 1$ or $b < 1$) can lead to extreme behavior in the posterior, such as a mode at the boundary (0 or 1) when the observed counts are very small. This may cause numerical instability or overly concentrated estimates. To ensure well-behaved posterior estimates and avoid such issues, we recommend using $a, b \geq 1$, which corresponds to a non-informative or

Table II. Important symbols used in Sections 2 and 3.

π	Probability of the Bernoulli distribution
$\hat{\pi}$	Maximum likelihood estimator of π
$\tilde{\pi}$	Bayesian estimator of π
λ	Regularization parameter
I	Fisher information
S	Sign test statistic
Z	Homogeneity test statistic
T	McNemar test statistic

weakly informative prior that provides shrinkage without producing extreme modes.

For all the three choices, the prior distribution is beta and the posterior distribution is also beta [Gelman et al. 2013]. The resulting estimator of π is obtained as the maximum of the posterior distribution and this may be approximated by $\hat{\pi}$ for all three choices for $n \rightarrow \infty$. Alternatively to taking the expectation of the posterior distribution, one can also use the mode, resulting in the maximum a posteriori (MAP) estimator, which tends to be more sensitive to the prior and may provide more stable estimates when data is sparse.

Intensive attention has been paid to the choice of a suitable prior in the Bayesian analysis of categorical data [Agresti 2002]. In [Tuyl et al. 2008], the focus was on estimating π for data with extreme outcomes and particularly with zero counts, i.e. with $X = 0$; naturally, $X = n$ represents an analogous complication. The conclusion was to avoid $a < 1$ and $b < 1$ for beta priors and thus to avoid Jeffreys prior. Among informative priors for the situation with some prior knowledge about π , the beta prior represents a very flexible and highly recommendable choice. The estimates $\tilde{\pi}_{BL}$ and $\tilde{\pi}_J$ are obtained for non-informative prior. We can say that the concept of “non-informative” prior is quite misleading, because the non-informative estimates $\tilde{\pi}_{BL}$ and $\tilde{\pi}_J$ do not correspond to the MLE.

Table II summarizes some important notation used throughout Sections 2 and 3, including symbols for probabilities, counts, and regularized estimates. For clarity, we explicitly note that hats denote maximum likelihood estimates, tildes denote Bayesian or shrinkage estimates, and stars denote λ -regularized statistics. This convention is used consistently throughout the paper to keep the notation transparent.

2.1. Properties of the Bayesian estimators

The following properties of the estimators of Table I may be derived in a straightforward way. Theorem 2 reveals the shrinkage estimators (1) to represent a compromise between MLE and $E\pi$. Theorem 3 evaluates the maximum possible

shrinkage for $\tilde{\pi}_{BL}$ and $\tilde{\pi}_J$. Theorem 4 compares the shrinkage intensity between $\tilde{\pi}_{BL}$ and $\tilde{\pi}_J$.

PROPERTY 1. It holds that $\tilde{\pi}_\beta = \hat{\pi} + \mathcal{O}\left(\frac{1}{n}\right)$ for $n \rightarrow \infty$.

PROPERTY 2. If $X_1, \dots, X_n$ are i.i.d. random variables following the Bernoulli distribution with parameter π , then it holds for $\gamma \geq 0$ and $\delta > 0$ that

$$\min \left\{ \frac{\sum_{i=1}^n X_i}{n}, \frac{\gamma}{\delta} \right\} \leq \frac{\sum_{i=1}^n X_i + \gamma}{n + \delta} \leq \max \left\{ \frac{\sum_{i=1}^n X_i}{n}, \frac{\gamma}{\delta} \right\}. \quad (2)$$

PROPERTY 3. If $\hat{\pi} < 1/2$, then

$$\max_{\hat{\pi} \in [0, 1/2)} |\hat{\pi} - \tilde{\pi}_{BL}| = \frac{1}{n+2} \quad (3)$$

and this value is attained either for $\hat{\pi} = 0$ or for $\hat{\pi} = 1$. If again $\hat{\pi} < 1/2$, then

$$\max_{\hat{\pi} \in [0, 1/2)} |\hat{\pi} - \tilde{\pi}_J| = \frac{1}{n+1} \quad (4)$$

and this value is attained either for $\hat{\pi} = 0$ or for $\hat{\pi} = 1$.

PROPERTY 4. It holds that

$$\tilde{\pi}_{BL} < \tilde{\pi}_J \iff \hat{\pi} > \frac{1}{2} \quad (5)$$

and

$$\tilde{\pi}_{BL} = \tilde{\pi}_J \iff \hat{\pi} = \frac{1}{2}. \quad (6)$$

It already follows that

$$\tilde{\pi}_{BL} > \tilde{\pi}_J \iff \hat{\pi} < \frac{1}{2}. \quad (7)$$

3. HYPOTHESIS TESTS BASED ON BAYESIAN ESTIMATES

Further, we are interested in the properties of common hypothesis tests for categorical data used with regularized versions of maximum likelihood estimates of the probabilities. Particularly, we consider the sign test for binomial distribution, Pearson χ^2 test of homogeneity, and McNemar test for 2×2 contingency tables. The regularized (shrunk) estimates of the probabilities of Section 2 will be considered in a natural way to obtain regularized versions of the test statistics. The asymptotic distribution of each of the test statistics is derived under the null hypothesis.

3.1. Sign test

We study the sign test for the binomial distribution assuming $X \sim \text{Bi}(n, \pi)$. We consider the null hypothesis $H_0 : \pi = 1/2$. The maximum likelihood estimator of π is denoted as $\hat{\pi} = X/n$.

Under the null hypothesis, the number of positive differences S_e exactly follows the binomial distribution, i.e. it holds $S_e \sim \text{Bi}(n, 0.5)$, where $\text{Bi}(n, 0.5)$ denotes the binomial distribution with n trials and success probability 0.5. In this paper, we will focus on the asymptotic version of the test, using the normalized statistic denoted as S . The asymptotic test is performed according to

$$S = 2\sqrt{n} \left(\hat{\pi} - \frac{1}{2} \right) = \frac{2X - n}{\sqrt{n}} \xrightarrow[H_0]{\mathcal{D}} \Omega, \quad \text{where } \Omega \sim N(0, 1). \quad (8)$$

This test statistic can be easily derived as the statistic $S = (\hat{p} - E\hat{p})/\sqrt{\text{var } \hat{p}}$, where the expectation under H_0 is $E\hat{p} = 1/2$ and the variance under H_0 is $\text{var } \hat{p} = 1/(4n)$.

THEOREM 1. Let us assume a random variable X following the binomial distribution $\text{Bi}(n, \pi)$. If the test statistic S (8) exploits the estimate $\pi^* = \lambda X + (1 - \lambda)\pi_0$ instead of $\hat{\pi}$, then the resulting test statistic S denoted as S^* fulfills

$$\frac{S^*}{\lambda} \xrightarrow[H_0]{\mathcal{D}} \Omega \quad (9)$$

under H_0 , where Ω follows $N(0, 1)$ distribution.

The considered penalized estimator $\pi^* = \lambda X + (1 - \lambda)\pi_0$ is a Bayesian estimate from Section 2 with a particular choice of $\lambda \in [0, 1]$.

3.2. Test of homogeneity

Consider a 2×2 contingency table, where each of the two columns is assumed to follow a binomial distribution. The Pearson χ^2 test of homogeneity will then be applied to assess whether the distributions across the columns are significantly different.

Let us assume the random samples to be measured in the total number of 2 populations (groups). A binary variable is observed in each randomly selected unit. The observed counts for the two groups are presented in the contingency table

	Group A	Group B	Σ
Success	n_{11}	n_{12}	$n_{1\cdot}$
Failure	n_{21}	n_{22}	$n_{2\cdot}$
Σ	$n_{\cdot 1}$	$n_{\cdot 2}$	n

(10)

Each column follows a binomial model with parameters (n_j, p_j) , where p_j is the probability of success in population j . The probabilities are presented in the table

	Group A	Group B
Success	π_1	π_2
Failure	$1 - \pi_1$	$1 - \pi_2$
Σ	1	1

(11)

The maximum likelihood estimates of π_1 and π_2 will be denoted by $\hat{\pi}_1$ and $\hat{\pi}_2$, respectively.

We consider the null hypothesis $H_0 : \pi_1 = \pi_2$. One of alternative ways of describing the χ^2 test for a 2×2 table is to consider the test statistic

$$Z = (\hat{\pi}_2 - \hat{\pi}_1) \sqrt{\frac{nn_{\cdot 1}n_{\cdot 2}}{n_{\cdot 1}n_{\cdot 2}}} \xrightarrow[H_0]{\mathcal{D}} \Xi, \quad (12)$$

where $\xrightarrow[H_0]{\mathcal{D}}$ denotes convergence of distribution and Ξ follows $N(0, 1)$ distribution. It holds that Z^2 is exactly equal to the χ^2 statistic of Pearson test; unlike χ^2 , the test statistic Z may also be used for a one-sided test.

To justify the test statistic Z , it is sufficient to show that the square of Z fulfils

$$\begin{aligned} Z^2 &= \left(\frac{n_{11}}{n_{\cdot 1}} - \frac{n_{12}}{n_{\cdot 2}} \right)^2 \frac{nn_{\cdot 1}n_{\cdot 2}}{n_{\cdot 1}n_{\cdot 2}} \\ &= \left(\frac{n_{11}n_{22} - n_{12}n_{21}}{n_{\cdot 1}n_{\cdot 2}} \right)^2 \frac{nn_{\cdot 1}n_{\cdot 2}}{n_{\cdot 1}n_{\cdot 2}} \\ &= \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{\cdot 1}n_{\cdot 2}n_{\cdot 1}n_{\cdot 2}} = \chi^2, \end{aligned} \quad (13)$$

i.e. Z^2 is equal to the Pearson χ^2 test statistic.

THEOREM 2. We consider a 2×2 contingency table. If the test statistic Z is considered with estimates

$$\pi_1^* = \lambda \frac{n_{11}}{n_{\cdot 1}} + (1 - \lambda) \frac{n_{1\cdot}}{n} \quad \text{and} \quad \pi_2^* = \lambda \frac{n_{12}}{n_{\cdot 2}} + (1 - \lambda) \frac{n_{1\cdot}}{n}, \quad (14)$$

then the resulting test statistic Z denoted as Z^* fulfills

$$\frac{Z^*}{\lambda} = \frac{(\pi_2^* - \pi_1^*)}{\lambda} \sqrt{\frac{nn_{\cdot 1}n_{\cdot 2}}{n_{\cdot 1}n_{\cdot 2}}} \xrightarrow[H_0]{\mathcal{D}} \Xi \quad (15)$$

under H_0 , where Ξ follows $N(0, 1)$ distribution.

The regularized estimates in (14) are Bayesian estimates from Section 2 with a particular choice of $\lambda \in [0, 1]$. While regularization is typically considered in the context of zero counts in contingency tables, the regularization is justified by the

connection to Bayesian thinking here.

We note that the shrinkage target for each group's success probability is taken as the pooled success rate, which is estimated from the data. This corresponds to an empirical Bayes approach. Importantly, the asymptotic standard normal limits derived remain valid under this choice.

3.3. McNemar test

Let us now consider a 2×2 table following a multinomial model with a fixed total number of samples n . McNemar test is a test of symmetry and at the same time a test of marginal homogeneity [Smith and Ruxton 2020]. Often, the test is applied if comparing the effect of a treatment before some event and after it. The table of observed counts will be denoted by

Before	After		Σ
	A	B	
A	n_{11}	n_{12}	$n_{1\cdot}$
B	n_{21}	n_{22}	$n_{2\cdot}$
Σ	$n_{\cdot 1}$	$n_{\cdot 2}$	n

(16)

The table of the corresponding probabilities has the form

Before	After		Σ
	A	B	
A	π_{11}	π_{12}	$\pi_{1\cdot}$
B	π_{21}	π_{22}	$\pi_{2\cdot}$
Σ	$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	1

(17)

The maximum likelihood estimates of the probabilities are denoted as $\hat{\pi}_{ij} = n_{ij}/n$ for $i = 1, 2$ and $j = 1, 2$. The test of $H_0: \pi_{12} = \pi_{21}$ is based on the statistic

$$T = \frac{n}{\sqrt{n_{12} + n_{21}}} (\pi_{12} - \pi_{21}) \xrightarrow[H_0]{\mathcal{D}} \Xi, \quad \text{where } \Xi \sim N(0, 1). \quad (18)$$

It can be shown easily that the square of T is equal to the commonly used form of the statistic χ^2 of McNemar test, which fulfils

$$\chi^2 = \frac{n_{12} - n_{21}}{n_{12} + n_{21}} \xrightarrow[H_0]{\mathcal{D}} \Xi, \quad \text{where } \Xi \sim \chi_1^2. \quad (19)$$

Let us now consider regularized estimators

$$\pi_{12}^* = \lambda \hat{\pi}_{12} + (1 - \lambda)\tau \quad \text{and} \quad \pi_{21}^* = \lambda \hat{\pi}_{21} + (1 - \lambda)\tau \quad (20)$$

for a given $\tau \in [0, 1]$ and $\lambda > 0$. The value of τ may be given by prior knowledge and the estimates in (20) correspond to Bayesian estimates presented in Section 2.

THEOREM 3. We consider a 2×2 contingency table. If the test statistic Z is considered with estimates (14), then the resulting test statistic, denoted as T^* , satisfies

$$T^* = \frac{n}{\sqrt{n_{12} + n_{21}}} \frac{(\pi_{12}^* - \pi_{21}^*)}{\lambda} \xrightarrow[H_0]{\mathcal{D}} \Xi, \quad \text{where } \Xi \sim N(0, 1). \quad (21)$$

The results of this section are intuitive, but analogous results do not seem to be available for larger contingency tables. Additionally, analogous reasoning for continuous data leads to much more complicated outcomes. For example, the two-sample t -test based on regularized means can be derived to follow a noncentral t -distribution, which presents a more complex situation with less straightforward application due to challenges in estimating the noncentrality parameter.

Given the simplicity of the test statistics, a practical approach for practitioners is to apply a parametric bootstrap using the shrunk probabilities. This allows assessment of finite-sample type I error rates and can improve the reliability of hypothesis testing when sample sizes are small.

Practitioners should be aware that the asymptotic standard normal limits for S^*/λ , Z^*/λ , and T^*/λ under H_0 require applying the scaling by λ when comparing to standard normal critical values. If the unscaled statistics S^* , Z^* , or T^* are used directly, the critical values must be adjusted accordingly to account for the factor λ .

4. REGULARIZED VERSIONS OF ASSOCIATION MEASURES

Using regularized test statistics, it is possible to obtain alternative versions of various association measures or other characteristics for categorical data. Let us assume the situation of Section 3 with a 2×2 contingency table. To avoid any misunderstanding, the regularized version of the statistic Z^* (37) will now be denoted as $Z^*(\lambda)$; the value of λ is always chosen according to the choices of Table I. Table III summarizes some important notation used throughout Sections 4 and 5.

THEOREM 4. Let us assume a 2×2 contingency table with the test statistic (12) denoted as Z_1 . Let us assume another 2×2 contingency table with the test statistic (12) denoted as Z_2 . Assuming a given $\lambda > 0$, let regularized versions of Z_1 and Z_2 be denoted as $Z_1^*(\lambda)$ and $Z_2^*(\lambda)$, respectively. If it holds $Z_1 > Z_2$, then it holds that $Z_1^*(\lambda) > Z_2^*(\lambda)$.

Table III. Important symbols used in Sections 4 and 5.

C_P	Pearson contingency coefficient
φ	Phi coefficient
V	Cramér's coefficient
MI	Mutual information
t_{ij}	Shrinkage target for $i = 1, \dots, I$ and $j = 1, \dots, J$
τ_i	Shrinkage target for $i = 1, \dots, I$
η_j	Shrinkage target for $j = 1, \dots, J$

Theorem 4 states that a regularized version of (12) is a suitable basis for regularized association measures. The proof follows from Theorem 2.

The regularized test statistic $Z^*(\lambda)$ preserves the main invariance properties of the original statistic. In particular, it is invariant to relabeling of categories and remains monotone in the odds ratio for 2×2 contingency tables. Moreover, while the regularization introduces a slight bias in small samples, this is accompanied by a reduction in variance, in line with the classical bias–variance tradeoff.

Let us now consider two commonly used measures of association for contingency tables derived directly from Z (12), or in fact from its square, which is the Pearson's χ^2 statistic. The measures are the Pearson contingency coefficient defined as

$$C_P = \sqrt{\frac{Z^2}{Z^2 + n}}, \quad (22)$$

and the phi coefficient defined as

$$\varphi = \sqrt{\frac{Z^2}{n}}. \quad (23)$$

THEOREM 5. Theorem 4 holds if the regularized statistic $Z^*(\lambda)$ (12) is replaced by

(1) the regularized Pearson contingency coefficient defined as

$$C_P^*(\lambda) = \sqrt{\frac{(Z^*(\lambda))^2}{(Z^*(\lambda))^2 + n}}, \quad (24)$$

(2) the regularized phi coefficient defined as

$$\varphi^*(\lambda) = \sqrt{\frac{(Z^*(\lambda))^2}{n}}. \quad (25)$$

For an overview of properties of the association measures, we refer to [Agresti 2002]. Another commonly used association measure is Cramér's coefficient V . We

may define its regularized version for an $I \times J$ table in the form

$$V = \sqrt{\frac{Z^2}{n(q-1)}}, \quad \text{where } q = \min\{I, J\}; \quad (26)$$

nevertheless, for a 2×2 table, $V^*(\lambda)$ coincides with the regularized phi coefficient (25).

5. REGULARIZED MUTUAL INFORMATION

In this section, we introduce a formal definition of a regularized version of mutual information for a two-way contingency table of size $I \times J$. While various regularized versions of mutual information have been employed in the analysis of real-world data (e.g. [Kalina and Schlenker 2015]), we aim to provide a more rigorous treatment here. Although the underlying idea is not novel, we believe that offering a more formal and theoretically precise definition is desirable. Our approach using the Bayesian estimates of Section 2 is inspired by [Hausser and Strimmer 2013], who formulated a regularized version of the Shannon entropy, where the regularized (shrinkage) version is known as Bayesian entropy. In Section 5.1, we derive that the regularized mutual information does not depend on cells with zero counts.

We consider a discrete variable X with categories $\{1, \dots, I\}$ and a discrete variable Y with categories $\{1, \dots, J\}$. Let us denote the joint probability of the category i for X and of the category j for Y by π_{ij} . In this notation, the parameters are $\pi_{11}, \dots, \pi_{IJ}$ and it holds naturally that $\sum_{i=1}^I \sum_{j=1}^J \pi_{ij} = 1$. The mutual information between the two variables X and Y is defined by

$$MI(X, Y) = - \sum_{i=1}^I \sum_{j=1}^J \hat{\pi}_{ij} \log \frac{\hat{\pi}_{ij}}{\hat{\pi}_{i \cdot} \hat{\pi}_{\cdot j}}. \quad (27)$$

Let $\hat{\pi}_{11}, \dots, \hat{\pi}_{IJ}$ denote maximum likelihood estimates of $\pi_{11}, \dots, \pi_{IJ}$. Let us consider regularized estimates of π_{ij} obtained as

$$\pi_{ij}^* = \lambda \hat{\pi}_{ij} + (1 - \lambda) t_{ij} \quad (28)$$

with a given $t_{ij} \in (0, 1)$ for all possible i and j , where $\lambda \in [0, 1]$ is a regularization parameter. We assume the shrinkage target $(t_{11}, \dots, t_{IJ})^T \in \mathbb{R}^{IJ}$ to be a vector of non-negative values. We assume also the vectors $\zeta = (\zeta_1, \dots, \zeta_I)^T \in \mathbb{R}^I$ and $\eta = (\eta_1, \dots, \eta_J)^T \in \mathbb{R}^J$ to be vectors of non-negative values. We assume

$$\sum_{i=1}^I t_{ij} = \eta_j \quad \text{for } j = 1, \dots, J, \quad \sum_{j=1}^J t_{ij} = \zeta_i \quad \text{for } i = 1, \dots, I, \quad (29)$$

and $\sum_{i=1}^I \sum_{j=1}^J t_{ij} = 1$.

Using the regularized estimates (28) within the population mutual information (27), we obtain the regularized empirical version of $\text{MI}(X, Y)$ formally defined by

$$\begin{aligned} \widehat{\text{MI}}(X, Y) = & - \sum_{i=1}^I \sum_{j=1}^J (\lambda \hat{\pi}_{ij} + (1 - \lambda) t_{ij}) \\ & \times [\log(\lambda \hat{\pi}_{ij} + (1 - \lambda) t_{ij}) - \log(\lambda \hat{\pi}_i + (1 - \lambda) \zeta_i) - \log(\lambda \hat{\pi}_j + (1 - \lambda) \eta_j)]. \end{aligned} \quad (30)$$

We note that the regularized mutual information remains nonnegative and equals zero if and only if X and Y are independent, provided that the vectors t , ζ , and η used for shrinkage are strictly positive and appropriately normalized. This ensures that the regularized MI preserves the fundamental properties of the classical mutual information, giving practitioners confidence in its use.

5.1. Zero counts

We investigate the contribution of zero counts to the value of the regularized mutual information. Zero counts are a common phenomenon in contingency tables with a large number of rows and/or columns. We show that the mutual information does not actually depend on these zero counts, which may therefore be disregarded in the formula (30).

THEOREM 6. We consider a discrete variable X with categories $\{1, \dots, I\}$ and a discrete variable Y with categories $\{1, \dots, J\}$ with the notation of Section 5. Let $\hat{\pi}_i > 0$ for every $i = 1, \dots, I$ and let $\hat{\pi}_j > 0$ for every $j = 1, \dots, J$. Let us have vectors of non-negative values $t \in \mathbb{R}^{IJ}$, $\zeta \in \mathbb{R}^I$, and $\eta \in \mathbb{R}^J$.

- (1) Let us consider a given pair $[i, j]$, for which $\hat{\pi}_{ij} = 0$. Let us assume $t_{ij} = 0$. Then it holds that

$$\begin{aligned} & (\lambda \hat{\pi}_{ij} + (1 - \lambda) t_{ij}) [\log(\lambda \hat{\pi}_{ij} + (1 - \lambda) t_{ij}) \\ & - \log(\lambda \hat{\pi}_i + (1 - \lambda) \zeta_i) - \log(\lambda \hat{\pi}_j + (1 - \lambda) \eta_j)] = 0. \end{aligned} \quad (31)$$

- (2) Let the set of all pairs $[i, j]$, for which $\hat{\pi}_{ij} = 0$, be denoted by $\mathcal{S}$. Then the regularized empirical mutual information fulfils

$$\begin{aligned} \widehat{\text{MI}}(X, Y) = & - \sum_{\substack{i=1, j=1 \\ [i, j] \notin \mathcal{S}}}^{IJ} (\lambda \hat{\pi}_{ij} + (1 - \lambda) t_{ij}) [\log(\lambda \hat{\pi}_{ij} + (1 - \lambda) t_{ij}) \\ & - \log(\lambda \hat{\pi}_i + (1 - \lambda) \zeta_i) - \log(\lambda \hat{\pi}_j + (1 - \lambda) \eta_j)]. \end{aligned} \quad (32)$$

We note that Theorem 6 assumes strictly positive row and column marginals. In practice, if an entire row or column is zero (which can occur in sparse or high-dimensional contingency tables), the corresponding categories can be removed from the computation of the regularized mutual information, or a small positive value can be imputed to ensure numerical stability. Such adjustments maintain the validity of the regularized MI while avoiding undefined logarithms.

6. CONCLUSION

This paper contributes to the growing body of work at the intersection of Bayesian estimation and categorical data analysis, offering valuable insights for addressing challenges commonly encountered in machine learning applications [Kalina and Rensová 2015]. By introducing regularized estimators and a novel uncertainty coefficient for measuring associations between categorical variables, we provide tools that can improve the robustness and interpretability of machine learning models, especially when dealing with high-dimensional data [Kalina and Matonoha 2020]. These methods offer promising directions for future research, particularly in areas such as feature selection, model calibration, and handling small or imbalanced datasets [Zhang et al. 2024]. We hope this work serves as a source of inspiration for researchers looking to advance statistical methodologies within the machine learning community, opening new avenues for improving model performance and reliability in real-world applications.

Important regularized approaches for the analysis of categorical data can be naturally obtained within the Bayesian framework, which allows combining observed data with prior information. Even in the absence of prior knowledge, it may be advantageous to consider Bayesian (shrinkage) estimates instead of relying solely on maximum likelihood estimates. A further benefit is that Bayesian methods provide the entire posterior distribution of parameters rather than just point estimates, enabling richer uncertainty quantification.

The effect of regularization on hypothesis tests and association measures can be complex to evaluate, with the exception of a few simple models discussed in this paper. In particular, when using regularized probabilities, the asymptotic null distribution of test statistics is modified, and the impact can be assessed in specific cases. Practical guidance on the choice of the regularization parameter λ is limited; however, if the variance of the prior distribution were known or could be reliably estimated, λ could be chosen to reflect the relative weight of prior information versus

data, providing a principled approach for practical applications. Developing systematic rules for selecting λ in this way remains an interesting direction for future research.

Future work will focus on developing novel hypothesis tests for approximated neural networks for categorical data with large numbers of parameters, particularly emphasizing high-dimensional contingency tables. These tests will serve as diagnostic tools, such as zero-weight tests, prioritizing interpretability and computational efficiency. Moreover, the insights presented in this paper offer opportunities to enhance machine learning models, enabling them to better handle large, sparse, and complex datasets.

APPENDIX: PROOFS

PROOF [of Theorem 1] The result follows from a simple reformulation

$$\tilde{\pi}_\beta = \hat{\pi} \frac{n}{n+a+b} + \frac{a}{n+a+b} = \hat{\pi} \left(1 - \frac{a+b}{n+a+b} \right) + \frac{a}{n+a+b} \quad (33)$$

and subsequently from

$$\tilde{\pi}_\beta = \hat{\pi} \left(1 - \mathcal{O} \left(\frac{1}{n} \right) \right) + \mathcal{O} \left(\frac{1}{n} \right), \quad n \rightarrow \infty. \quad (34)$$

PROOF [of Theorem 2] The result follow immediately from the following property. Let $\alpha \geq 0$, $\beta > 0$, $\gamma \geq 0$, and $\delta > 0$. Then it holds that

$$\min \left\{ \frac{\alpha}{\beta}, \frac{\gamma}{\delta} \right\} \leq \frac{\alpha + \gamma}{\beta + \delta} \leq \max \left\{ \frac{\alpha}{\beta}, \frac{\gamma}{\delta} \right\}. \quad (35)$$

PROOF [of Theorem 1] We may express

$$\begin{aligned} \frac{1}{\lambda} S^* &= \frac{1}{\lambda} 2\sqrt{n} \left(\pi^* - \frac{1}{2} \right) \\ &= \frac{1}{\lambda} 2\sqrt{n} \left(\lambda \frac{X}{n} + (1-\lambda) \frac{1}{2} - \frac{1}{2} \right) \\ &= \frac{1}{\lambda} 2\sqrt{n} \left(\lambda \frac{X}{n} - \lambda \frac{1}{2} \right) \\ &= S, \end{aligned} \quad (36)$$

from which the asymptotic distribution of (8) immediately follows.

PROOF [of Theorem 2] The result follows from

$$Z^* = \left(\lambda \frac{n_{12}}{n_2} - \lambda \frac{n_{11}}{n_1} \right) \sqrt{\frac{nn_{.1}n_{.2}}{n_1 n_2}} = \lambda Z. \quad (37)$$

PROOF [of Theorem 3] The result follows from

$$\begin{aligned} T^* &= \frac{n}{\sqrt{n_{12} + n_{21}}} (\lambda \hat{\pi}_{12} + (1 - \lambda) \tau - \lambda \hat{\pi}_{21} - (1 - \lambda) \tau) \\ &= \frac{n}{\sqrt{n_{12} + n_{21}}} (\lambda \hat{\pi}_{12} - \lambda \hat{\pi}_{21}) = \lambda T. \end{aligned} \quad (38)$$

ACKNOWLEDGEMENTS

The author would like to thank an anonymous referee for valuable suggestions and Pavel Rak and Anežka Faltýnková (both MFF UK) for technical help.

REFERENCES

- AGRESTI, A. 2002. *Categorical Data Analysis*, 2nd ed. Wiley, Hoboken.
- BAEK, J. AND PARK, J. 2023. Mixture of networks for clustering categorical data: A penalized composite likelihood approach. *American Statistician* 77, 259–273.
- BAYES, T. 1763. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London* 53, 370–418.
- BRAUN, F., CAELEN, O., SMIRNOV, E., KELK, S., AND LEBICHOT, B. 2017. Improving card fraud detection through suspicious pattern discovery. In *Lecture Notes in Computer Science*. Vol. 10351. Springer, Cham, 181–190.
- DE TOLEDO, P., NÚÑEZ, F., AND USABIAGA, C. 2020. Matching in segmented labor markets: An analytical proposal based on high-dimensional contingency tables. *Economic Modelling* 93, 175–186.
- DENG, J. AND DENG, Y. 2022. Maximum entropy of random permutation set. *Soft Computing* 26, 11265–11275.
- GELMAN, A., CARLIN, J., STERN, H., DUNSON, D., VEHTARI, A., AND RUBIN, D. 2013. *Bayesian Data Analysis*, 3rd ed. Chapman and Hall/CRC, Boca Raton.
- GOLDEN, R. 2020. *Statistical Machine Learning: A Unified Framework*. Chapman & Hall/CRC Press, Boca Raton.
- GUPTA, R., GUPTA, S., SINGH, J., AND KAIS, S. 2025. Entropy-assisted quality pattern identification in finance. *Entropy* 27, 4, 430.
- HAUSSER, J. AND STRIMMER, K. 2013. Entropy inference and the james-stein estimator with application to nonlinear gene association networks. *Journal of Machine Learning Research* 10, 1469–1484.
- ISSOUANI, E., BERTAIL, P., AND GAUTHERAT, E. 2024. Exponential bounds for regularized hotelling’s t^2 statistic in high dimension. *Journal of Multivariate Analysis* 203, 105342.
- JAYNES, E. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge.
- JOHNSON, A., OTT, M., AND DOGUCU, M. 2022. *Bayes Rules! An Introduction to Applied Bayesian Modeling*. CRC Press, Boca Raton.
- KALINA, J. AND MATONOKA, C. 2020. A sparse pair-preserving centroid-based supervised learning method for high-dimensional biomedical data or images. *Biocybernetics and Biomedical Engineering* 40, 2, 774–786.

-
- KALINA, J. AND RENSOVÁ, D. 2015. How to reduce dimensionality of data: Robustness point of view. *Serbian Journal of Management* 10, 131–140.
- KALINA, J. AND SCHLENKER, A. 2015. A robust supervised variable selection for noisy high-dimensional data. *BioMed Research International* 2015, 320385.
- KALINA, J. AND TICHAVSKÝ, J. 2022. The minimum weighted covariance determinant estimator for high-dimensional data. *Advances in Data Analysis and Classification* 16, 977–999.
- LEDOIT, O. AND WOLF, M. 2022. The power of (non-)linear shrinking: A review and guide to covariance matrix estimation. *Journal of Financial Econometrics* 20, 187–218.
- LINDSKOU, M., ERIKSEN, P., AND TVEDEBRINK, T. 2020. Outlier detection in contingency tables using decomposable graphical models. *Scandinavian Journal of Statistics* 47, 347–360.
- LOFTUS, S. 2024. *An Introductory Handbook of Bayesian Thinking*. Elsevier, London.
- MING, H. AND YANG, H. 2024. l_0 regularized logistic regression for large-scale data. *Pattern Recognition* 146, 110024.
- POSE, F., BAUTISTA, L., GIANMUSO, F., AND REDELICO, F. 2021. On the permutation entropy bayesian estimation. *Communications in Nonlinear Science and Numerical Simulation* 99, 105779.
- RAO, C. 2002. *Linear Statistical Inference and Its Applications*. Wiley, New York.
- SMITH, M. AND RUXTON, G. 2020. Effective use of the mcnemar test. *Behavioral Ecology and Sociobiology* 74, 133.
- SOHAEI, N. 2023. Error and optimism bias regularization. *Journal of Big Data* 10, 8.
- SUBRAMANIAN, I., VERMA, S., KUMAR, S., JERE, A., AND ANAMIKA, K. 2020. Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights* 2020, 14.
- TUYL, F., GERLACH, R., AND Mengersen, K. 2008. A comparison of bayes-laplace, jeffreys, and other priors: The case of zero events. *American Statistician* 62, 40–44.
- WANG, A., HENAO, R., AND CARIN, L. 2024. Transformer in-context learning for categorical data. arXiv:2405.17248.
- WANG, K., LI, J., AND TSUNG, F. 2023. Efficient and interpretable monitoring of high-dimensional categorical processes. *IIE Transactions* 55, 886–900.
- WANG, R. AND LI, J. 2023. Block-regularized 5×2 cross-validated mcnemar’s test for comparing two classification algorithms. Submitted.
- ZHANG, Y., ZAIDI, N., ZHOU, J., WANG, T., AND LI, G. 2024. Effective interpretable learning for large-scale categorical data. *Data Mining and Knowledge Discovery* 38, 2223–2251.
- ZHOU, X., HENG, Q., CHI, E. C., AND ZHOU, H. 2024. Proximal mcmc for bayesian inference of constrained and regularized estimation. *The American Statistician* 78, 4, 379–390.

Jan Kalina

The Czech Academy of Sciences, Institute of Computer Science, Prague
& Charles University, Faculty of Mathematics and Physics, Prague

E-mail: kalina@cs.cas.cz